Speech Rhythm and the Three Aspects of Speech Timing: Articulatory, Acoustic and Auditory*

Ilsung Yun**

ABSTRACT

This study is targeted at introducing the three aspects of speech timing (articulatory, acoustic and auditory) and discussing their strong and weak points in describing speech timing. Traditional (extrinsic) articulatory timing theories exclude timing representation in the speaker's articulatory plan for his utterance, while the (intrinsic) articulatory timing theories headed by Fowler incorporate time into the plan for an utterance. As compared with articulatory timing studies with crucial constraints in data collection, acoustic timing studies can deal with even several hours of speech relatively easily. This enables us to perform suprasegmental timing studies as well as segmental timing studies. On the other hand, perception of speech timing is related to psychology rather than physiology and physics. Therefore, auditory timing studies contribute to enhancing our understanding of speech timing from the psychological point of view. Traditionally, some theories of speech timing (e.g. typology of speech rhythm: stress-timing; syllable-timing or mora-timing) have been based on our perception. However, it is problematic that auditory timing can be subjective despite some validity. Many questions as to speech timing are expected to be answered more objectively. Acoustic and articulatory description of timing will be the method of solving such problems of auditory timing.

Keywords: speech rhythm (timing), articulatory, acoustic, auditory, typology

1. Introduction

It has been believed that every language has its own typical rhythm or timing pattern, which is regarded as one of the important factors in making a language sound different from others. There seems to be a general consensus about the definition of rhythm, i.e., alternation and succession of the same sequences (Classe, 1939; Allen, 1975). On the other hand, we use another term, timing. The concept of timing, however, does not appear to be well established, despite its frequent use. Therefore, the two terms often do not seem to be distinguished from each other. However, it is felt likely that the term timing has been used in a wider sense than rhythm, i.e., to describe overall temporal phenomena. This would be

---

* A summary of this paper was presented at the 17th Conference of Speech Communication and Signal Processing organized by the Acoustical Society of Korea (August 25-26, 2000).

** Division of Liberal Arts (English), University of Seoul.

due to the constraint that the definition of rhythm has. Speech must involve various nonrhythmic temporal phenomena as well as rhythmicality, for the description of which the term timing would be more appropriate.

Speech exists in three different forms (i.e., articulatory, acoustic and auditory). Therefore, rhythm (timing) in speech also could be described in the following three dimensions:

(1) **articulatory timing** in the production of speech
(2) **acoustic timing** in the process of transmission of speech from speaker to hearer
(3) **auditory timing** from the hearer's point of view.

In this paper, we will introduce each of the three aspects of speech timing through a literature review and discuss their strong or weak points to describe rhythm (timing) in speech.

2. Articulatory Timing

In the field of phonetics, the issue of how to explain coarticulation has long been a matter of debate. As is well known, no sound (phoneme) is not coarticulated, unless it is produced in isolation. Therefore, the account of coarticulation is directly related to the understanding of articulation. With respect to the study of articulatory timing, the paper written by Fowler (1980) is regarded as a turning point, in which a new timing theory (i.e., action theory) is introduced. In that paper, Fowler criticised the previous studies regarding coarticulation (e.g., Kozhevnikov & Chistovich, 1965; Öhman, 1966, 1967; MacNeilage, 1970; Kent & Netsell, 1971; Daniloff & Hammarberg, 1973; Kent, Carney & Severeid, 1974; Kent & Moll, 1975; Hammarberg, 1976; Kent & Minifie, 1977; Bell-Berti & Harris, 1979), naming them *extrinsic timing theories*, and proposed the *intrinsic timing theory* as an alternative. Even before Fowler (1980), there had been indications that none of the current theories gave adequate explanations of coarticulation. For instance, Kent & Minifie (1977), through a literature review, pointed out that none of the models and theories of coarticulation suggested until then explained the phenomenon to the full. The variety in the models and theories themselves may indirectly imply some lack of explanatory power in them. In such a situation, what Fowler (1980) contributed to the study of timing would be the classification of the previous studies into a group and the radical suggestion of a new theory.

On the other hand, most of the theories and models for coarticulation mentioned in Fowler (1980) were developed independently, and furthermore, the authors did not use the term *extrinsic timing* in their papers directly (even if the expression *extrinsic allophones* is

seen in Kent & Minifie, 1977:123). Therefore, it is difficult to classify them into a group, as Fowler herself acknowledges. However, Fowler claims that the traditional theories all share the same view: "they exclude timing from representation in the talker's articulatory plan for his utterance. Instead, they propose that an utterance is given coherence in time only by its actualization." (p. 113).

Fowler's definition of the extrinsic timing theory is clearly exposed in the following remark: "In an extrinsic timing theory, the articulatory plan represents the serial ordering of features, segments or syllables - i.e., it represents their ordinal relationships along the time axis - but time is not taken to inhere in, or to be essential to, the specification of these production units." (p. 113).

The criticism against extrinsic timing is focused on the assumptions the so-called extrinsic timing theorists share: (1) the essential properties of segments are timeless, (2) segments are temporally discrete, and (3) the plan of an utterance is distinct from its executor (Fowler, 1980:116). Against these assumptions, Fowler insisted that (1) "there are no borders perpendicular to the time axis in an articulatory or acoustic record to separate one segment from another" (p. 114), and the successive segments could be distinguished only qualitatively (2) a new theory (intrinsic timing) "must merge the plan and its executor by incorporating time into the plan for an utterance." (p. 122). Fowler also disagreed with the suggestion that phonological segments are subjective (Hammarberg, 1976), which was based on the subjective (optical) and objective (physical) perception (motion) developed by some psychologists (e.g., Gibson, 1957, 1968, 1969, 1977; Polanyi, 1958). That is, between the two possibilities that segments are perceived either objectively (in acoustic signal) or subjectively (with no acoustic support), Hammarberg supported the latter. But Fowler (1980:121) expressed a different view: "it [phonological segment] is neither subjective nor objective, but something else that spans the dichotomy."

The action theory (intrinsic timing) is based on the notion of coordinative structures which were so named in Easton (1972). On the assumption that all our behaviours are coordinated, Fowler (1980) suggests its three significant properties. First, the coordinative structure is an organization that spans several muscles. Second, a small coordinative structure is nested in a larger structure. Third, many coordinative structures are cyclic (e.g., walking, chewing, breathing). Using the notion of coordinative structures, the intrinsic timing theory accounts for coarticulation: "The plan for an utterance is treated as identical to the coordinative structures themselves." (p. 128), and defines it as the coproduction (overlapping production) of successive, continuous, four dimensional (i.e., static/dynamic, and discrete/continuous) *canonical* segments. On the heels of Fowler (1980), other studies, e.g., Bell-Berti & Harris (1981), Tuller, Harris and Kelso (1982), and Harris, Tuller & Kelso (1986) appeared. They observed some signs of invariance in the timing of articulatory events in a VCV sequence despite the variations of stress and speech rate. On the basis of the invariance, it

was claimed that articulatory events are controlled by a temporally invariant coordinative structure.

On the other hand, the action theory has been criticised, in particular with regard to perception, since a theory of production requires a corresponding theory of perception. Those who criticised the action theory include Diehl, 1986; Fujimura, 1986; Ohala, 1986; Massaro, 1986; Remez, 1986; Porter, 1986; Studdert-Kennedy, 1986; Lindblom & MacNeilage, 1986; Lindblom, 1987. Their criticisms are, for example, as follows: (1) what we perceive is not the articulatory source but the speech sound (Ohala, Massaro), (2) articulated phonetic segments have no ecological significance and perceivers are not aware of it (Diehl, Remez), (3) articulated phonetic segments are not autonomous events (objects of perception) to perceivers (e.g., Ohala: a glottal stop /ʔ/ has a different significance for English speakers and Arabic speakers respectively; Studdert-Kennedy; Remez), (4) the intentionality to make an utterance cannot be explained in the action theory (intrinsic timing) (Lindblom & MacNeilage), (5) speech production is versatile, i.e., variant (Lindblom), etc. Despite these criticisms, the action theory must be seen as having provided the theory of timing (particularly articulatory timing) with a new view.

3. Acoustic Timing

Putting aside the above criticisms about either extrinsic timing or intrinsic timing, articulatory timing studies seem to have possessed a crucial limit in their data collection. In other words, it has been difficult for their data (speech materials) to be larger units than a phoneme or a syllable, because detecting the temporal invariance in the articulatory movements (interactions of muscles) across a long stretch of speech is difficult. Though in recent years, optical tracking systems and electromagnetic measuring devices as well as X-ray micro-beam procedures can provide a solution to the problem, some equipment (e.g., X-ray micro-beam) is not widely available and could be even dangerous. Even if all technical problems are solved, it is still pointed out that the movements of articulatory organs (muscles) are very complicated and overlap with each other. Therefore, to see a flourish of articulatory timing studies (especially intrinsic timing studies), we may have to wait for the development of some radical techniques. In contrast, acoustic timing studies are not under such constraints. That is, the acoustic speech signals in a larger speech unit (even several hours of speech) can be analysed relatively easily owing to the rapid development of acoustic instruments. Furthermore, conventional linguistic or rhythmic units are reasonably well defined, and there exists general agreement about the segmentation of acoustic signals (Fletcher, 1988). This situation enables us to perform suprasegmental timing studies as well as segmental timing studies. The numerous acoustic data bases which have been developed

in speech research laboratories can contribute a great deal to suprasegmental timing studies (e.g., stress, rhythm, pause, tone, intonation, and emotional speech).

In addition, the development of speech synthesis has been aided by acoustic timing studies. Klatt (1976, 1979) in American English and O'Shaughnessy (1981, 1984) in French, etc. would be the prototype studies in speech synthesis. This type of study (acoustic or peripheral timing studies) does not consider the articulatory (central) timing mechanisms such as muscle movements. Instead, they almost exclusively deal with "syntagmatic ordering of phoneme-type segments" (Fletcher, 1988:11).

Most of the studies of the so-called objective speech rhythm or timing, which criticise the classic type of rhythm studies, are based on the acoustic data of speech. Given that a large-scale study of articulatory timing is impractical, an acoustic timing study would be chosen as an objective study more frequently. In addition, even if articulatory timing studies using a long sequence of connected speech become easily available in the future, the physical identity of speech would remain our primary concern in studies of speech timing or rhythm, as eventually what we hear is not the articulatory movements but the speech sound.

4. Auditory Timing

It can be said that speech is completed when it is finally perceived by a listener. But the problem is that the ear, (or to be more exact, our perception) is not like a scale or ruler. In other words, our auditory mechanism (from the eardrum to the brain) does not perceive the speech sound (physical characteristics: intensity, frequency, duration) as is. Therefore, what we hear (perceive) could significantly differ from the physically-existing speech. In line with this, Denes & Pinson (1993:98) state that "The difference between the physical properties of a sound and the subjective properties of the same sound cannot be stressed too strongly." Due to this difference, perceptual timing of speech has been considered independent of articulatory and acoustic timing.

The studies of time perception were pioneered by psychologists (e.g., Blakely, 1933; Woodrow, 1951; Fraisse, 1963), as part of rhythm theory. They mentioned the so-called *indifference interval*, which is noteworthy with reference to speech timing as well. According to Woodrow (1951), the fact that short intervals are overestimated and long ones are underestimated was first stated by Vierordt (1868) as a law. Woodrow (1951:1225) interpreted Vierordt's statement as implying that "some intermediate length is neither overestimated nor underestimated", and he stated "this intermediate length is the indifference interval." The indifference interval is said to range from 0.6 to 0.8 s (Blakely, 1933; Woodrow, 1951; Fraisse, 1963), in spite of the denial that it is too variant to be determined

(Treisman, 1963; Michon, 1967). This concept has been applied to speech research by some scholars including Allen (1975) and Lehiste (1977). Accepting the concept of the indifference interval, Allen (1975:77) claimed that "The over- and underestimation of relatively short and long intervals ... may explain our ability to hear sequences of approximately equal time intervals as more equal than they really are." Through a series of investigations into the so-called just noticeable differences in duration, Lehiste (1977) reported that they ranged from 30 to more than 100 ms, and predicted that the just noticeable differences for speech would be somewhat larger, since listeners noticed the durational change better with nonsense stimuli (e.g., pure tones, white noise) than they did with speech. Based on the results and Allen's view mentioned above, Lehiste supported the perception of regular rhythm (isochronism) in languages with stress accent.

While perceptual timing studies contribute to enhancing our understanding of some aspects of timing in speech, we have to acknowledge that they have restrictions in themselves. That is, perception of timing could be subjective, which may cause significant variability between speakers and within speakers. Yet our questions as to timing are expected to be answered more objectively.

In relation to the subjectivity of perception of speech rhythm, we come across an interesting report. In Miller's (1984) study, four groups of subjects (11 English phoneticians, 27 English non-phoneticians, 11 French phoneticians and 10 French non-phoneticians) were exposed to seven language recordings in reading and conversational styles (Arabic, Polish, Spanish, Finnish, Japanese, Indonesian and Yoruba). All groups indicated that Arabic was stress-timed. Interestingly, however, Spanish was judged stress-timed by the three groups except the group of English non-phoneticians. This is a result contrary to the view of Pike (1945) and Hockett (1955) who judged Spanish as syllable-timed. In addition, none of the remaining five languages was perceived to be identical in rhythm by all four groups, while some languages were not assigned to either rhythmic category. For example, Polish was felt to be strongly stress-timed by English phoneticians, whereas it was judged strongly syllable-timed by English non-phoneticians. None of the four groups could assign Indonesian or Japanese to either rhythmic category. Therefore, it will be unrealistic to classify the rhythmic type of languages only depending on perception, even though some notable agreement is likely to be obtained (e.g., Arabic or Spanish?).

On the other hand, perceptual timing could be influenced by the phonology of the listener's native or familiar language. With reference to this, Jakobson, Fant & Halle (1952: 10) asserted that the perception of speech is significantly affected by the phonology of our native language.

> ....., the way we perceive them [the speech sounds] is determined by the phonemic pattern familiar to us. Therefore, a monolingual Slovak identifies the rounded front

> vowel /ø/ of the French word *jeu* as /e/, since the only distinctive opposition in his mother tongue is acute (front) *vs.* grave (back) and not flat (rounded) *vs.* plain (unrounded). A monolingual Russian, on the contrary, perceives the same French vowel as /o/ because his native tongue possesses only the one of the two oppositions in question, namely, flat *vs.* plain.

Jakobson, Fant & Halle (1952:10-11) went on to make another interesting statement that this interference by phonology could take place even in the responses to non-speech sounds.

> Knocks produced at even intervals, with every third louder, are perceived as groups of three separated by a pause. The pause is usually claimed by a Czech to fall before the louder knock, by a Frenchman to fall after the louder; while a Pole hears the pause one knock after the louder. The different perceptions correspond exactly to the position of the word stress in the languages involved: in Czech the stress is on the initial syllable, in French, on the final and in Polish, on the penult. When the knocks are produced with equal loudness but with a longer interval after every third, the Czech attributes greater loudness to the first knock, the Pole, to the second, and the Frenchman, to the third.

It seems to be true that we hear foreign languages through our learned sound system (phonology). There are two other examples of interference by phonology in our perception. First, Allen & Ladefoged (1971) examined the effect of differences in segmental structure between English, French and Polish on perceived rhythm. Nonsense utterances (VC#CV, V#CCV and V#CV) spoken by native English, French and Polish speakers were judged by other native speakers. That is, the listeners tapped their fingers whenever they heard the stressed syllables. The results indicated that "(1) listeners hear the rhythm of an utterance more clearly when it is spoken by a compatriot than by a foreigner, and (2) listeners tend to perceive rhythmic structure in terms of their own linguistic rules, regardless of the language of the utterance" (p. 116). Second, Berinstein (1978) performed a cross-linguistic perception test to determine how listeners' language background influences stress perception. The subjects of the experiment were speakers of three different languages with different stress patterns respectively (i.e., English: variable; Spanish: penultimate; K'ekchi: fixed final). It was found that to the same sequences of four synthetic CV syllables (in each sequence three of four syllables were equal duration (100 ms), while the duration of one remaining syllable varied between 70 and 200 ms), the listeners showed systematically different responses. That is, an initial bias was shown in English, no bias in Spanish, and a final bias was shown in K'ekchi (Berinstein suggested no clear explanation for the

reason why an initial bias was shown in English while no bias in Spanish, but it could be noted that the responses systematically varied according to the native language of the subjects). With regard to the result, Berinstein (1978:S56) claimed that "the response bias pattern for the three groups is in part governed by the linguistic structure and the stress assignment rules in a particular language." All in all, we could not rule out the possibility that perception of timing can be affected, to some extent, by the phonology of the listener's native language.

5. Conclusion

The rhythm theory of speech has been based mainly on perceptual timing. Particularly in the past, the theorists of speech rhythm mainly depended on their impression (perception) of speech. The major instruments that they would use were their ears or fingers for tapping. Despite some validity, their theory has left many questions unanswered. Later, perceptual isochrony was proposed especially for the explanation of stress-timing in English. But it also is not likely to be the ultimate solution that we have waited for. Accordingly, so as to secure objectivity and generalisation in the theory of speech timing, we have to go beyond the limits imposed by perceptual timing. Considering the above literature review and discussion, it is suggested that studies of acoustic timing or articulatory timing are a key to solving the problem.

References

Allen, G. D. 1975. Speech rhythm: its relation to performance universals and articulatory timing, *Journal of Phonetics*, **3**, 75-86.

Allen, G. D. & Ladefoged, P. 1971. Syllable Structure and Sentence Rhythm - A Cross Language Study, *Journal of the Acoustical Society of America*, **50**, 116.

Bell-Berti, F. & Harris, K. 1979. Anticipating coarticulation: Some implications from a study of liprounding, *Journal of the Acoustical Society of America*, **65**, 1268-1270.

Bell-Berti, F. & Harris, K. 1981. A temporal model of speech production, *Phonetica*, **38**, 9-20.

Berinstein, A. 1978. A cross-linguistic study on the perception of stress, *Journal of the Acoustical Society of America*, **63**, S 55-56.

Blakely, W. 1933. *The discrimination of short empty temporal interval*, Ph.D. Dissertation, University of Illinois.

Classe, A. 1939. *The Rhythm of English Prose*, Oxford: Blackwell.

Daniloff, R. G. & Hammarberg, R. E. 1973. On defining coarticulation, *Journal of Phonetics* **1**, 239-248.

Denes, P. B. & Pinson, E. N. 1993. *The Physics and Biology of Spoken Language*, 2nd ed.,

W. H. Freeman and Company.

Diehl, R. 1986. Coproduction and direct perception of phonetic segments: a critique, *Journal of Phonetics*, **14**, 61-66.

Easton, T. 1972. On the normal use of reflexes, *American Scientist* **60**, 591-599.

Fletcher, J. 1988. *An Acoustic Study of Timing in French*, Ph. D. thesis (unpublished), University of Reading.

Fowler, C. A. 1980. Coarticulation and theories of extrinsic timing, *Journal of Phonetics*, **8**, 113-133.

Fraisse, P. 1963. *The Psychology of Time*, New York: Harper and Row.

Fujimura, O. 1986. Evaluating the Task Dynamic model, *Journal of Phonetics*, **14**, 105-108.

Gibson, J. J. 1957. Optical motions and transformations as stimuli for visual perception, *Psychological Review* **64**, 288-295.

Gibson, J. J. 1968. What gives rise to the perception of motion?, *Psychological Review* **75**, 335-346.

Gibson, J. J. 1969. The change from visible to invisible: A study of optical transitions, *Perception & Psychophysics* **5**, 113-116.

Gibson, J. J. 1977. The theory of affordances. In *Perceiving, acting and knowing: Toward an Ecological Psychology*. (R. Shaw & J. Bransford, eds) 67-82, Hillsdale, NJ: Erlbaum.

Hammarberg, R. 1976. The metaphysics of coarticulation, *Journal of Phonetics* **4**, 353- 363.

Harris, K. S., Tuller, B. and Kelso, J. A. S. 1986. Temporal invariance in the production of speech, In Perkell, J. & Klatt, D. H. (eds.) *Invariance and variability in speech processes*, Hillsdale, N. J.: Lawrence Erlbaum Associates, 243-252.

Hockett, C. 1955. *A manual of phonology*, Memoir 11, Indiana University Publication in Anthropology and Linguistics, Bloomington.

Jakobson, R., Fant, G. & Halle, M. 1952. *Preliminaries to Speech Analysis; the Distinctive Features and Their Correlates*, Cambridge, Massachusetts: the M.I.T.

Kent, R. D., Carney, P. J. & Severeid, L. R. 1974. Velar movement and timing: Evaluation of a model for binary control, *Journal of Speech and Hearing Research* **17**, 470-488.

Kent, R. D. & Minifie, F. D. 1977. Coarticulation in recent speech production models, *Journal of Phonetics* **5**, 115-117.

Kent, R. D. & Moll, K. L. 1975. Articulatory timing in selected consonant sequences, *Brain and Language* **2**, 304-323.

Kent, R. D. & Netsell, R. 1971. Effects of stress contrast on certain articulatory parameters, *Phonetica* **24**, 23-44.

Klatt, D. H. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence, *Journal of the Acoustical Society of America*, **59**, No. 5, 1208- 1221.

Klatt, D. H. 1979. Synthesis by rule of segmental durations in English sentences, In Lindblom, B. & Öhman, S. (eds.) *Frontiers of Speech Communication Research*, New York: Academic Press, 287-299.

Kozhevnikov & Chistovitch 1965. *Speech: Articulation and Perception*, Joint publications Research Service 30, 543. U.S. Department of Commerce, Washington D. C.

Ladefoged, P. 1993. *A Course in Phonetics*, New York, Harcourt Brace Jovanovich, Inc. 3rd ed.

Lehiste, I. 1977 Isochrony reconsidered, *Journal of Phonetics*, **5**, 253-264.

Lindblom, B. E. F. 1987. Adaptive variability and absolute constancy in speech signals: two themes in the quest for phonetic invariance, *Proceedings of the Eleventh Inter-*

national Congress of Phonetic Sciences, Tallinn, Estonia, USSR **3**, 9-18.
Lindblom, B. E. F. & MacNeilage, P. 1986. Action theory: problems and alternative approaches, *Journal of Phonetics* **14**, 117-132.
MacNeilage, P. 1970. Motor control of serial ordering of speech, *Psychological Review* **77**, 182-196.
Massaro, D. W. 1986. A new perspective and old problems, *Journal of Phonetics*, **14**, 69-74.
Michon, J. A. 1967. *Timing in Temporal Tracking*, Institute for Perception RVO- TNO. Soesterberg, The Netherlands.
Miller, M. 1984. On the perception of rhythm, *Journal of Phonetics*, **12**, 75-83.
Ohala, J. 1986. Against the direct realist view of speech production, *Journal of Phonetics* **14**, 75-82.
Öhman, S. 1966. Coarticulation in VCV utterance: Spectrographic measurements, *Journal of the Acoustical Society of America*, **39**, 151-168.
Öhman, S. 1967. Numerical models of coarticulation, *Journal of the Acoustical Society of America*, **41**, 310-320.
O'Shaugnessy, D. 1981. A study of French vowel and consonant durations, *Journal of Phonetics* **9**, 385-406.
O'Shaugnessy, D. 1984. A multispeaker analysis of durations in read French paragraphs, *Journal of the Acoustical Society of America*, **76**, 1664-1672.
Pike, K. 1945. *Intonation of American English*. University of Michigan Press: Ann Arbor.
Polanyi, M. 1958. *Personal Knowledge*. Chicago: University of Chicago Press.
Porter, R. J. 1986. Speech messages, modulations and motions, *Journal of Phonetics* **14**, 83-88.
Remez, R. E. 1986. Realism, language and other barriers, *Journal of Phonetics* **14**, 89- 97.
Studdert-Kennedy, M. (1986) Two cheers for direct realism, *Journal of Phonetics* **14**, 99-104.
Treisman, M. 1963. Temporal discrimination and the indifference interval, *Psychological Monographs* **77**, 1-31.
Tuller, B., Harris, K. S. & Kelso, J. A. S. 1982. Stress and rate: differential transformation of articulation, *Journal of the Acoustical Society of America* **71**, 1534-1543.
Vierordt, K. 1868. *Der Zeitsinn nach Versuchen*, Tübingen: H. Laupp.
Woodrow, H. 1951. Time Perception, *Handbook of Experimental Psychology*, (Stevens, S. S. Ed.), New York: Wiley, 1224-1236.

Received: Jan. 30, 2001.
Accepted: Feb. 25, 2001.

▲ ILSUNG YUN
Division of Liberal Arts - English
University of Seoul
90 Junnong-dong Dongdaemun-gu
Seoul 130-743, KOREA
Tel: +82-2-2210-2458
E-mail: highland@kebi.com